Objective Evaluation of Phase and Magnitude Only Reconstructed Speech: New Considerations

Erfan Loveimi, *Student Member, IEEE*, and Seyed Mohammad Ahadi, *Senior Member, IEEE*

Speech Processing Research Laboratory
Electrical Engineering Department, Amirkabir University of Technology, Hafez Ave., Tehran 15914, Iran
{eloveimi, sma}@aut.ac.ir

ABSTARCT

The aim of this paper is to make some improvements in the synthesis part of Analysis-Modification-Synthesis (AMS) framework to reconstruct speech with higher quality. In previous works based on this framework [7-12] the number of iterations was fixed on 100. This number of iterations was not chosen in accordance with a specific measure. In this paper, we monitored the quality of reconstructed speech per iteration and through a heuristic criterion, controlled the number of iterations. Our results showed that the required number of iterations depends on frame duration, applied window and whether the speech signal is being reconstructed from magnitude or phase spectrum. In case of phase-only speech reconstruction we need larger number of iterations. By applying rectangular window, the phase-only reconstructed speech surpasses its magnitude-only counterpart in frames longer than 64 ms, qualitatively. However, in case of using Hamming window, depending on applying LSEE or OAM, phase-only reconstructed speech will have better quality than its magnitude-only counterpart in frames longer than 256 ms and 128 ms, respectively.

Index Terms- *Iterative speech reconstruction, phase spectrum, magnitude spectrum, quality assessment.*

1. INTRODUCTION

There is established a general belief that the phase spectrum of speech signal plays negligible role in speech processing. By taking a glance on the main areas of research in the speech processing community, we can only see the trace of phase in speech coding. In primary algorithms of speech coding [1], in order to reduce the bit-rate, speech was coded from magnitude spectrum. At the decoding stage, phase spectrum was constructed from magnitude spectrum and speech signal was reconstructed by combining them. However, the quality of reconstructed speech was not satisfactory. In fact such algorithms suffer from a structural problem. Reconstructing phase spectrum from magnitude spectrum involves particular assumptions to be held, i.e. the signal should be minimum phase [2]. However, speech signal is not. Consequently, the researchers tend to utilize phase spectrum information in speech coding process.

In other fields of speech processing, such as speech enhancement or speech recognition, almost all of the process is focused on the magnitude spectrum. For example, in speech enhancement methods, such as wiener filtering [3] or spectral subtraction [4], there is no role for phase to play. Most of the speech recognition front-ends, such as MFCC or LPC, attempt to extract the features from only the magnitude spectrum. Although some features which are based on group delay have been proposed in [5], they are not widespread yet.

The first comprehensive study about the importance of phase spectrum in signal processing was conducted by Oppenheim and Lim [6]. They studied the importance of phase in some types of signals such as speech and image. In case of speech signal, they observed that when frames become larger than 1 sec, phase spectrum-based reconstructed speech will be intelligible. However, this point had no remarkable influence on speech research since most of the operations in speech processing were carried out on frames with duration of 20-40 ms, which was due to speech non-stationarity.

Liu *et al* [7] carried out the first significant experiments to evaluate the importance of phase in speech recognition. They decomposed speech, which were stop consonants in intervocalic context, into frame durations from 16 to 512 ms, with 50% overlap along with Hamming windowing. Then, they reconstructed the signals from their magnitude and phase-only spectra. After generating phase and magnitude-only stimuli, they played them back to a number of listeners in order to recognize them. The reported recognition results showed that the phase-only reconstructed speech, in frame durations longer than 128 ms, surpasses its magnitude-only counterpart and becomes more intelligible.

Alsteris and Paliwal have conducted a series of remarkable experiments in recent years [8], [9]. Their framework, like Liu *et al*, was based on Analysis-Modification-Synthesis (AMS). They showed that phase spectrum, even in short frame duration such as 32 ms, could have a remarkable deal of intelligibility information in case of using rectangular window. The readers are referred to [9] for more profound study about their researches.

As speech is not a minimum phase signal, the phase spectrum cannot be constructed through Hilbert transform from magnitude spectrum. As a result the signal could not be reconstructed in time domain uniquely. Hence, we should seek for an iterative signal reconstruction procedure in order to reconstruct the speech from its phase-only or magnitude-only spectra. In this case, we should monitor the relationship between the number of iterations and specific objective or subjective quality measures. Whenever the change of the measure entered a specific range or tolerance, the iterative algorithm should be stopped. The next iterations only have a computational cost with no notable quality improvement. In the previous related tasks [7-12] the number of iterations was set to 100. In the meantime, [10] and [12] reported that almost there is no remarkable change in the quality of the reconstructed speech after 30 iterations. In [10], the authors used subjective tests to evaluate the quality of speech. In this case, it is very difficult to precisely monitor the speech quality change per iteration. Hence there remains a question: what is the appropriate number of iterations for reconstructing the speech signal from its magnitude or phase-only spectra? It is clear that this depends on the situation i.e. frame duration and window type. In this paper, we want to investigate the relationship among the aforementioned factors. Our pilot studies showed that although the quality of magnitude-only reconstructed speech will not change after almost 35 iterations, in case of phase-only speech reconstruction this number of iterations is not enough.

We reconstructed the speech signal from its magnitude-only and phase-only spectra via Least Square Error Estimation (LSEE) [12] and Overlap-Add Method (OAM) in different frame lengths. The effects of rectangular and Hamming windows have been studied too. In order to evaluate the speech quality, PESQ [13] objective measure has been used.

In section 2 of this paper we will briefly review the AMS framework. Section 3 will discuss the quality assessment process. In section 4, the problem will be stated in more details. Finally, the experiments, their results and analysis will be presented in section 5.

2. ANALYSIS-MODIFICATION-SYNTHESIS FRAMEWORK

Speech is a quasi-stationary signal. Hence it cannot be analyzed directly by Fourier transform. First, it should be decomposed into frames in which the stationarity assumption could be held. After applying a window, *w(n)*, it is analyzed by taking Fourier transform. The result is called Short-Time Fourier Transform (STFT) [14]:

$$X_n(\omega) = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)e^{-j\omega m}. \qquad (1)$$

where *x(n)* is speech signal, $X_n(\omega)$ is short-time Fourier transform and *n* in (1) denotes the index of STFT. Due to the complexity of $X_n(\omega)$, it can be decomposed in the following manner:

$$X_n(\omega) = |X_n(\omega)|e^{j\psi_n(\omega)}, \qquad (2)$$

where $|X_n(\omega)|$ and $\psi_n(\omega)$ denote short-time magnitude and phase spectra of *n*th frame, respectively.

Next step is modification. It is performed to investigate the significance of specific parameter(s). For example, to evaluate the importance of phase spectrum in different situations such as different frame durations and window types, one can reconstruct the signal from its phase-only spectrum. The same can be done for magnitude spectrum.

Usually to reconstruct the signal from its phase spectrum, magnitude spectrum is replaced by a constant number. It is common to set magnitude spectrum to unity:

$$\hat{X}_n^1(\omega) = e^{j\psi_n(\omega)}, \qquad (3)$$

where $\hat{X}_n$ and $\psi_n(\omega)$ denote the modified spectrum and short-time phase spectrum of the *n*th frame. The superscript of 1 points to the initialization of iterative speech reconstruction algorithm.

In order to reconstruct the speech from its magnitude spectrum, one can replace the phase spectrum with a sequence of random uniformly distributed numbers in the range of $(-\pi, \pi)$ or $(0, 2\pi)$, like φ:

$$\hat{X}_n^1(\omega) = |X_n(\omega)|e^{j\varphi}. \qquad (4)$$

φ also can be set to zero. The results of our simulations do not show a significant difference, thus we initialized the magnitude-only speech reconstruction algorithm with zero phase, i.e. φ is set to zero.

The next and final step of this framework is synthesis. We have utilized two procedures, OAM and LSEE. OAM is a well-known method with the following formula:

$$x^{i+1}(n) = \frac{\sum_{p=-\infty}^{\infty} \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}_p^i(\omega)e^{j\omega n}d\omega}{\sum_{p=-\infty}^{\infty} w(pL-n)}, \qquad (5)$$

where $x^{i+1}(n)$ denotes reconstructed signal after *i*+1th iteration, *p* is the frame number, *L* is decimation factor and ω denotes frequency. Note that in the case of phase-only signal reconstruction, $\hat{X}_p^i(\omega) = |X_p^i(\omega)|\angle X_p(\omega)$ and it is initialized by (3). In the case of magnitude-only speech reconstruction, $\hat{X}_p^i(\omega) = |X_p(\omega)|\angle \hat{X}_p^i(\omega)$ and it is initialized by (4). Most of the previous works [7-9] were based on OAM.

LSEE [12] is an alternative for OAM. Its formula is as following:

$$x^{i+1}(n) = \frac{\sum_{p=-\infty}^{\infty} w(pL-n)\frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}_p^i(\omega)e^{j\omega n}d\omega}{\sum_{p=-\infty}^{\infty} w^2\,(pL-n)}. \, (6)$$

The description of (6) is similar to (5).

When the spectrum of a signal is modified, it does not necessarily remain a valid spectrum. Hence, there may be no signal with such a spectrum. LSEE tries to find a signal with most similar spectrum to the modified spectrum in sense of mean square error (MSE). In fact, this method tries to find a signal which minimizes the following distance measure between reconstructed speech spectrum (X_{Rec}) and modified speech spectrum (X_{Mod}):

$$D[X_{Mod}(p,\omega), X_{Rec}(p,\omega)] = \sum_{p=-\infty}^{\infty} \frac{1}{2\pi}\int_{\omega=-\pi}^{\pi} |X_{Mod}(p,\omega) - X_{Rec}(p,\omega)|^2 \, \mathrm{d}\omega. \qquad (7)$$

Griffin and Lim [12] showed that their proposed algorithm i.e. LSEE, decreases the following distance measure (D_M) in each iteration:

$$D_M[x(p,n), x^i(p,n)] = \sum_{p=-\infty}^{\infty} \frac{1}{2\pi}\int_{\omega=-\pi}^{\pi} [|X(p,\omega)| - |X^i(p,\omega)|]^2 \, \mathrm{d}\omega, \qquad (8)$$

where x^i , *p* and ω denote the reconstructed speech after *i*th iteration, frame number and frequency, respectively. Fig. 1 shows D_M versus number of iterations in case of magnitude-only speech reconstruction. As we can see both OAM and LSEE reconstructed speech utterances do not have significant change after about 35 iterations.

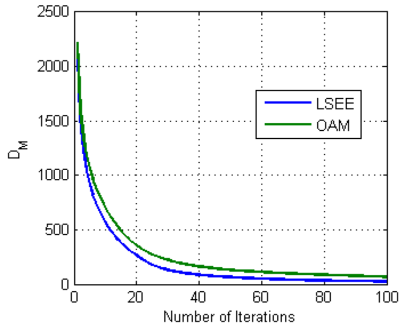


Figure 1. Distance measure versus number of iterations in case of magnitude-only speech reconstruction via OAM and LSEE.

3. QUALITY ASSESSMENT

Generally, there are two main approaches to evaluate the quality of speech signal: subjective and objective. Subjective methods are more reliable, but they suffer from two main problems. They are both time-consuming and costly. Objective measures were proposed to solve these problems. Although these measures can be easily and quickly calculated by computer, their reliability is questionable. The more correlation with the subjective measures, the better the objective measure. Hu and Loizou [15] carried out a comprehensive study on many types of objective measures to investigate their correlation with subjective tests. Their tests results found PESQ as a winner because of its higher correlation with subjective measures. Hence we used this objective measure to evaluate the quality of reconstructed speech.

PESQ was proposed by ITU-T to evaluate the quality of speech in telephone handsets and narrowband speech codecs [13]. It represents an aggregation of PAMS and PSQM99 [13]. PESQ produces a score between 1.0 and 4.5, with high values indicating better quality.

4. PROBLEM STATEMENT

The number of iterations for speech signal reconstruction in the previous studies [7-12] was fixed on 100. The authors of [10] and [12] stated that they did not observe a significant change after almost 30 iterations. Their claim was, to some extent, based on results similar to what we reported in Fig. 1. First of all, we should note that D_M is not a reliable measure to evaluate the level of similarities between two speech signals. In fact, it is too general for comparing the similarity of two signals and utilizes no specific information about their nature. In [10], subjective measures as well as D_M were used to evaluate the quality of reconstructed speech. Although subjective tests could be very reliable, they suffer from a few problems. In case of applying them, monitoring the quality change of the reconstructed speech per iteration becomes very difficult. Actually, human ear does not possess such a resolution to identify slight changes of speech quality per iteration. Here, we used PESQ objective measure to monitor the changes in the quality of reconstructed speech per iteration. This method is much more reliable than using D_M or subjective tests.

Fig. 2 shows the quality of the reconstructed signal from its phase spectrum. As seen clearly, 100 iterations are not sufficient and could not realize all the potentials of phase spectrum in reconstructing the speech signal. Here, we have used a particular approach to stop the iterative signal reconstruction algorithm. We defined a relative error, Err,:

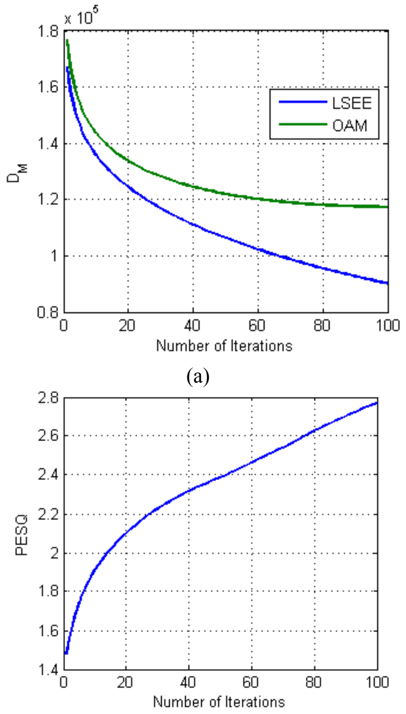
$$Err = \frac{|PESQ^{i-1} - PESQ^{i}|}{PESQ^{i-1}}, \qquad (8)$$

where i denotes the iteration number. If Err becomes less than 0.001, for 3 times, the iteration will be stopped.

5. EXPRIMENTS, SIMULATION RESULTS AND ANALYSIS

The utilized speech signals are chosen from NOIZEUS database [16]. This database is composed of 30 sentences with 8 kHz sampling rate and 16-bit precision. Here, we used 10 sentences in our experiments. Hence each point in figures and table represent the average of 10 results. The speech signals are decomposed into frame lengths of 32, 64, 128, 256, 512 and 1024 ms. Phase-only and magnitude-only stimuli were reconstructed through OAM and LSEE. The effects of Hamming and rectangular window are considered, too. The overlap and FFT size were set to 87.5% [8] and 2N, respectively. N is the number of samples of each frame.

Figs. 3 and 4 show the results of speech quality evaluation with PESQ objective measure. As we can see the crossover point in which the phase-only reconstructed speech surpasses its magnitude-only counterpart depends on frame duration, window type and synthesis method. Fig. 3 shows the results of applying Hamming window. In case of reconstructing the signal via LSEE, phase-only reconstructed speech surpasses its magnitude-only counterpart in frames longer than 256 ms, qualitatively. However, in case of using OAM the crossover point will shift backward and would be around 128 ms.



(a)

(b)

Figure 2. (a) D_M versus number of iterations, (b) PESQ versus number of iterations, in case of reconstructing the signal from its phase spectrum. As we can see this number of iterations (100) is not sufficient.

In comparison to our previous work [11], in which the number of iterations was fixed in 100, the crossover point was located in shorter frame durations due to the improvement of the quality of phase-only reconstructed speech.

LSEE leads to better quality in comparison to OAM in case of magnitude-only speech reconstruction. However, in case of phase-only speech reconstruction OAM will result in higher speech quality.

By applying rectangular window, the quality of phase-only reconstructed speech will be improved and the quality of magnitude-only reconstructed speech will be deteriorated. To reconstruct the signal from its magnitude or phase spectra we need a specific trade-off between the frequency resolution and frequency leakage of the window. In case of magnitude-only speech reconstruction, Hamming window provides better compromise. In case of phase-only speech reconstruction, rectangular window is the optimum choice.

As we can see in Fig. 4, the crossover point is located in frames longer than 64 ms. This is an interesting result which shows that phase spectrum even in short frame lengths such as 64 ms, could have a remarkable deal of intelligibility information, even more than magnitude.

The other point that should be mentioned here is the number iterations which meet the condition. The utilized stopping rule is not very critical and any similar one can be used. As we can see, the magnitude-only speech reconstruction algorithm will converge with more speed and requires lower number of iterations, while in case of using the phase spectrum to reconstruct the signal, further number of iterations is needed. Table I shows the average number of iterations which meet the stopping condition in each case.

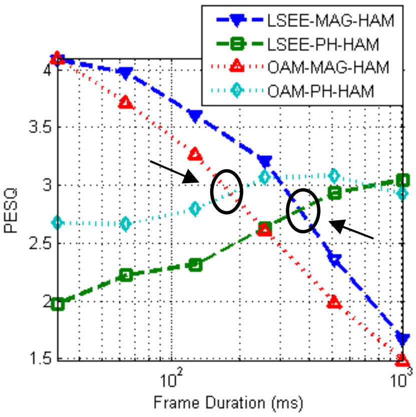


Figure 3. Phase-only reconstructed speech, along with Hamming window, surpasses its magnitude-only counterpart qualitatively in frames longer than 128 and 256 ms (denoted by black circles), in case of reconstructing the signal with OAM and LSEE, respectively.

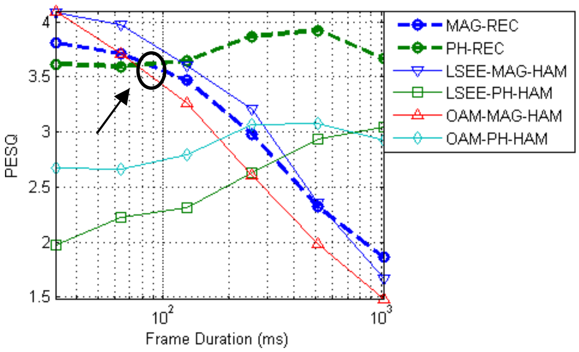


Figure 4. In case of using rectangular window both LSEE and OAM lead to the same results. As we can see, phase-only reconstructed speech surpasses its magnitude-only counterpart in frame longer than 64 ms (denoted by black circle).

Table 1. Average number of iterations that meet the stopping condition.

Frame Duration (ms)	Average of Number of Iterations					
	Magnitude			Phase		
	Rect*	Ham* (LSEE)	Ham (OAM)	Rect	Ham (LSEE)	Ham (OAM)
32	29.2	37.5	37.4	89.6	105.4	163.1
64	37.4	42.5	39.3	95.1	104.2	154.9
128	50.4	54.2	50.8	107.7	113.6	161.8
256	62.2	72.9	56.2	122.4	141.8	149.0
512	59.2	67.0	60.1	115.2	115.2	102.9
1024	56.3	56.5	66.2	84.5	85.6	56.3

Rect*: Rectangular window, Ham*: Hamming window.

*Generally, phase-only speech reconstruction algorithm in comparison with its magnitude-only counterpart requires more number of iterations to converge.

6. CONCLUSION

In this paper we made some modifications in AMS framework to improve the quality of the reconstructed speech. The number of iterations in similar tasks was fixed on 100. We showed that although magnitude-only speech reconstruction algorithm requires less number of iterations than 100, this number of iterations is not sufficient for phase-only speech reconstruction. To determine the appropriate number of iterations we used a heuristic criterion. We defined a relative error in accordance with objective measure change per iteration. We showed that the appropriate number of iterations depends on whether the signal is reconstructed from its magnitude or phase spectra, as well as the frame duration and applied window. In case of using rectangular window, the quality of phase-only reconstructed speech surpasses its magnitude-only counterpart in frames longer than 64 ms. However, in case of using Hamming window, the phase-only reconstructed speech quality surpasses that of the magnitude-only reconstructed speech in frames longer than 128 ms and 256 ms for OAM and LSEE, respectively.

ACKNOWLEDGMENT

This research was supported in part by the Iran Telecommunication Research Center (ITRC) under grant No.T-500-21804.

REFERENCES

[1] R.J. McAulay and T.F. Quatieri, *Sinusoidal coding, in Speech Coding and Synthesis*. New York: Elsevier, 1995, pp. 121–173.

[2] A.V. Oppenheim, R.W. Schafer, *Discrete-Time Signal Processing, second ed., Prentice* Hall, Upper Saddle River, NJ, 1999.

[3] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*. MIT Press, Cambridge, MA, 1949.

[4] S.F. Boll, .Suppression of Acoustic Noise in Speech using Spectral Subtraction, *IEEE Trans. on ASSP*, vol. 27(2), pp.113.120, 1979.

[5] P.S. Murthy, B. Yegnanarayana, Robustness of group-delay based method for extraction of significant instants of excitation from speech signals, *IEEE Trans. Speech Audio Process*. 7 (6) (1999) 609–619.

[6] A.V. Oppenheim, J.S. Lim, The importance of phase in signals, Proc. IEEE 69 (1981) 529–541.

[7] L. Liu, J. He, and G. Palm, "Effect of phase on the perception of intervocalic stop consonants," *Speech Commun.*, vol. 22, no. 4, pp. 403–417, 1997.

[8] K.K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eur. Conf. Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, Sep. 2003, pp. 2117–2120.

[9] L.D. Alsteris and K.K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital signal processing*, vol. 17, pp. 578–616, May 2007.

[10] P. Aarabi, G. Shi, M. Modir Shanechi, '*On the Importance of Phase in Human Speech Recognition', IEEE Trans. Acoust., Speech, Signal Process., vol. 14,* no. 5, Sep. 2006.

[11] E. Loveimi and S.M. Ahadi, "*Objective Evaluation of Magnitude and Phase Only Spectrum-based Reconstruction of the Speech Signal"*, in *Proc. Int. Symp.* On Communications, Control and Signal Processing (*ISCCSP 2010*), Limassol, Cyprus, Mar. 2010.

[12] *D.W. Griffin and J.S. Lim, "Signal estimation from modifi*ed short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[13] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, ITU-T Rec. P. 862, 2000.

[14] L. Rabiner, and B.H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[15] Y.Hu, and P.C.Loizou, "Evaluation of objective quality measures for speech enhancement*," IEEE Trans. Audio, Speech, Lang. Process.* 1*6, 229–238, 2008.*

[16] Y.Hu, *"NOIZEUS:* A noisy speech corpus for evaluation of speech enhancement algorithms," http://www.utdallas.edu/loizou/speech/noizeus/, 2005.